



## RESEARCH ARTICLE

# Generalized mean $p$ -values for combining dependent tests: comparison of generalized central limit theorem and robust risk analysis [version 1; peer review: 1 approved]

Daniel J. Wilson

Big Data Institute, Nuffield Department of Population Health, University of Oxford, Oxford, OX3 7LF, UK

**V1** First published: 31 Mar 2020, 5:55  
<https://doi.org/10.12688/wellcomeopenres.15761.1>  
 Latest published: 31 Mar 2020, 5:55  
<https://doi.org/10.12688/wellcomeopenres.15761.1>

## Abstract

The test statistics underpinning several methods for combining  $p$ -values are special cases of generalized mean  $p$ -value (GMP), including the minimum (Bonferroni procedure), harmonic mean and geometric mean. A key assumption influencing the practical performance of such methods concerns the dependence between  $p$ -values. Approaches that do not require specific knowledge of the dependence structure are practically convenient. Vovk and Wang derived significance thresholds for GMPs under the worst-case scenario of arbitrary dependence using results from Robust Risk Analysis (RRA).

Here I calculate significance thresholds and closed testing procedures using Generalized Central Limit Theorem (GCLT). GCLT formally assumes independence, but enjoys a degree of robustness to dependence. The GCLT thresholds are less stringent than RRA thresholds, with the disparity increasing as the exponent of the GMP ( $r$ ) increases. I motivate a model of  $p$ -value dependence based on a Wishart-Multivariate-Gamma distribution for the underlying log-likelihood ratios. In simulations under this model, the RRA thresholds produced tests that were usually less powerful than Bonferroni, while the GCLT thresholds produced tests more powerful than Bonferroni, for all  $r > -\infty$ . Above  $r > -1$ , the GCLT thresholds suffered pronounced false positive rates. Above  $r > -1/2$ , standard central limit theorem applied and the GCLT thresholds no longer possessed any useful robustness to dependence.

I consider the implications of these results in the context of various interpretations of GMPs, and conclude that the GCLT-based harmonic mean  $p$ -value procedure and Simes' (1986) test represent good compromises in power-robustness trade-off for combining dependent tests.

## Open Peer Review

### Reviewer Status

Invited Reviewers

1

#### version 1

31 Mar 2020



1. Raif Rustamov , AT&T Inc, Bedminster, USA

Any reports and responses or comments on the article can be found at the end of the article.

## Keywords

Combined tests, p-values, generalized means, generalized central limit theorem, robust risk analysis, harmonic mean p-value, dependent tests

**Corresponding author:** Daniel J. Wilson ([daniel.wilson@bdi.ox.ac.uk](mailto:daniel.wilson@bdi.ox.ac.uk))

**Author roles: Wilson DJ:** Investigation, Writing – Original Draft Preparation

**Competing interests:** No competing interests were disclosed.

**Grant information:** D.J.W. is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (Grant 101237). D.J.W. is supported by a Big Data Institute Robertson Fellowship.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2020 Wilson DJ. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Wilson DJ. **Generalized mean p-values for combining dependent tests: comparison of generalized central limit theorem and robust risk analysis [version 1; peer review: 1 approved]** Wellcome Open Research 2020, 5:55 <https://doi.org/10.12688/wellcomeopenres.15761.1>

**First published:** 31 Mar 2020, 5:55 <https://doi.org/10.12688/wellcomeopenres.15761.1>

## 1 Introduction

Combining  $p$ -values is a convenient and widely used form of meta-analysis that aggregates evidence across studies or tests, e.g. 1,2. Aggregating evidence in this way improves the sensitivity (power) of formal tests to detect subtle signals in data, making better use of resources and improving the potential for scientific discovery. There are many methods for combining tests, e.g. 3,4. In general, combining tests using the full data is more powerful than using summary statistics. However, access to the full data may be difficult for many reasons, for example computational tractability or issues of consent in statistical genetics. For these reasons, parameter estimates and standard errors, or Z-statistics, are often provided instead, e.g. 5. Combining Z-statistics, rather than  $p$ -values, allows parameters to be jointly estimated across datasets, e.g. 6. Nevertheless, combining  $p$ -values may be preferred when: (i) parameters are dataset-specific, (ii) hypotheses are mutually exclusive, or (iii) only the  $p$ -values are available. Fisher's is a widely-used method in scenario (i) that is appropriate when the datasets are independent<sup>7</sup>. The harmonic mean  $p$ -value (HMP) is suited to scenario (ii)<sup>8,9</sup>. The Bonferroni procedure<sup>10</sup> is a universal method for combining  $p$ -values under arbitrary dependence. These methods are closely connected and can be thought of as occupying different strategies in trading off power against robustness to dependence.

### Box 1. Kinds of combined tests for $p$ -values

Some methods for combining  $p$ -values were originally formulated as constructing simultaneous confidence intervals for multiple tests, including the Bonferroni and Šidák procedures<sup>10,11</sup>. They account for multiple comparisons by widening the confidence interval of each test from  $100(1 - \epsilon)\%$  to  $100(1 - \epsilon/K)\%$  and  $100(1 - \epsilon)^{1/K}\%$ , assuming arbitrary dependence and independence respectively, where  $K$  is the number of tests. A  $p$ -value can be defined as the widest  $100(1 - p)\%$  confidence interval that rejects the null hypothesis. Therefore, these approaches are equivalent to increasing the stringency of the  $p$ -value threshold of each test from  $\epsilon$  to  $\epsilon/K$  and  $1 - (1 - \epsilon)^{1/K}$  respectively. When any individual  $p$ -value falls below the adjusted threshold, the grand null hypothesis that none of the tests are significant can be rejected. Adjusting confidence intervals and significance thresholds for multiple testing is thus equivalent to a combined test in which the minimum  $p$ -value is compared against the adjusted significance threshold. For instance, Tippet's combined test<sup>12</sup>, in which the minimum  $p$ -value is compared to significance threshold  $1 - (1 - \epsilon)^{1/K}$  is equivalent to Šidák correction for this reason. In this article, I do not distinguish between these formulations of combining  $p$ -values. All aim to limit the probability of falsely rejecting the grand null hypothesis that none of the individual tests is significant, at a pre-determined level  $\epsilon$  (the weak-sense family-wise error rate<sup>13</sup>). Some combined tests (e.g. 9–11) are additionally able to reject subsets of tests when some null hypotheses are false while limiting the probability of falsely rejecting any true null hypotheses (the strong-sense family-wise error rate<sup>13,14</sup>).

The focus of this article is to consider significance thresholds for the generalized mean  $p$ -value (GMP) and compare their performance under different dependence assumptions. The test statistics underpinning the Bonferroni, Šidák, HMP, Fisher and other procedures are special cases of the GMP:

$$M_{r,K}(p_1, \dots, p_K) = \left( \frac{p_1^r + \dots + p_K^r}{K} \right)^{1/r} \quad (1)$$

which includes the maximum (when  $r \rightarrow \infty$ ), arithmetic mean ( $r = 1$ ), geometric mean ( $r \rightarrow 0$ ), harmonic mean ( $r = -1$ ) and minimum ( $r \rightarrow -\infty$ )<sup>15</sup>. The exponent parameter  $r$  affects the characteristics of the test, so that as  $r$  approaches  $-\infty$ , the GMP is more influenced by smaller  $p$ -values, and as it approaches  $\infty$  it is more influenced by larger  $p$ -values. These characteristics affect the interpretation of the GMP as suitable for particular purposes, such as model averaging (HMP;  $r = -1$ ), or combining evidence (Fisher's method;  $r \rightarrow 0$ ). (See section 6 for more on interpretation).

In general, GMPs cannot be interpreted directly as  $p$ -values because they are not uniformly distributed even when the null hypothesis (that the constituent  $p$ -values are uniformly distributed) is true<sup>15</sup>. Instead the GMP can be used as a test statistic by calculating a significance threshold  $\Psi_{r,K}(\epsilon)$  for rejecting the null hypothesis, which limits the false positive rate to some pre-specified level, e.g.  $\epsilon = 5\%$ :

$$\Pr \left( M_{r,K}(p_1, \dots, p_K) \leq \Psi_{r,K}(\epsilon) \mid \bigcap_{k=1}^K p_k \sim U(0,1) \right) \leq \epsilon. \quad (2)$$

This requires an assumption about dependence between the constituent  $p$ -values. For example, the  $p$ -values may be assumed (i) to be independent, as in Fisher's method<sup>7</sup>, (ii) to conform to a particular model of dependence, as in Brown's method<sup>16</sup>, or (iii) to possess arbitrary dependence, in which the worst case is usually considered, as in the Bonferroni procedure<sup>10</sup>.

Different assumptions about dependence produce different significance thresholds, which in turn affect the power of the test to reject the null hypothesis. More conservative assumptions produce more stringent thresholds that trade off reduced statistical power against greater robustness of the false positive rate to dependence. Vovk and Wang<sup>15</sup> derived significance thresholds for GMPs under arbitrary dependence by considering the worst case scenario using robust risk analysis (RRA). For the HMP they derived a significance threshold of

$$\Psi_{\text{RRA},-1,K}(\epsilon) = \frac{\epsilon}{\log K} \quad (3)$$

assuming large  $K$ . The result is precise (asymptotically as  $K \rightarrow \infty$ ) in the sense that for any value of  $\epsilon$ , there is a form of dependence under which the threshold is not conservative (so the equality in Equation 2 is satisfied)<sup>15</sup>.

In contrast, Wilson<sup>9</sup> derived a considerably less stringent threshold using generalized central limit theorem (GCLT):

$$\Psi_{\text{GCLT},-1,K}(\epsilon) = \frac{\epsilon_1}{1 + \epsilon_1 \log K}, \quad (4)$$

where  $\epsilon_1 \approx \epsilon$  and  $K$  is assumed large. This result implies that the HMP can be directly interpreted as if it were a  $p$ -value when it is small, because

$$\Psi_{\text{GCLT},-1,K}(\epsilon) \rightarrow \epsilon \quad \text{as} \quad \epsilon \rightarrow 0. \quad (5)$$

The difference in the stringency of the GCLT and RRA thresholds, which approaches  $\log K$  for small  $\epsilon$ , stems from different assumptions about dependence between the constituent  $p$ -values. Formally, the GCLT derivation assumes independence, but the heavy-tailed distribution of  $p^{-1}$  confers robustness to dependence<sup>9,17</sup>. Specifically, a result by Davis and Resnick<sup>18</sup> implies that Equation 5 holds despite dependence subject to the condition that

$$\Pr(p_j < \epsilon | p_i < \epsilon) \rightarrow 0 \quad \text{as} \quad \epsilon \rightarrow 0, \quad i \neq j, \quad i, j = 1, \dots, K. \quad (6)$$

However, Goeman, Rosenblatt and Nichols<sup>19</sup> reported simulations under a model of dependence satisfying the Davis-Resnick condition in which the GCLT threshold for  $\epsilon = 0.05$  incurred a false positive rate of 0.09. This raises several questions:

1. What forms of dependence are relevant when combining  $p$ -values?
2. Does the Davis-Resnick condition confer on the GCLT threshold adequate robustness to such dependence for practically relevant values of  $\epsilon$ , e.g. 0.05?
3. Do GMPs with exponents  $r \neq -1$  enjoy a more favourable power-robustness trade-off?

To address these questions, I derived significance thresholds for GMPs using GCLT (section 2). I motivated a model of  $p$ -value dependence based on the Wishart-Multivariate-Gamma distribution (section 3). I simulated under this model to test the power and false positive rates of the GCLT and RRA significance thresholds (section 4). To complete the picture, I derived procedures to control the strong-sense family-wise error rate based on the GCLT thresholds (section 5) and considered the interpretation of combining  $p$ -values using GMPs (section 6). The results indicate that the power of the GMP to combine  $p$ -values, under relevant dependence assumptions at  $\epsilon = 0.05$ , was better than the Bonferroni procedure for GCLT thresholds, but worse than the Bonferroni procedure for RRA thresholds. However, GCLT thresholds began to suffer pronounced false positive rates for  $r > -1$ , and enjoyed apparently no robustness to dependence whatever for  $r > -1/2$ . I conclude that the GCLT-based HMP procedure<sup>9</sup> and the related Simes (1986) test<sup>20</sup> represent good compromises in power-robustness trade-off for combining dependent  $p$ -values. These methods are interpretable in terms of model-averaging and require no specific knowledge of the dependence structure.

## 2 GCLT significance thresholds for generalized mean $p$ -values

This section uses GCLT to infer the distribution of GMPs under the grand null hypothesis and thereby construct significance thresholds, assuming the number of constituent  $p$ -values  $K$  is large. The GCLT derivation formally assumes the  $p$ -values are independent, but the Davis-Resnick condition extends to  $-\infty < r < 0$ , which implies that

robustness to dependence is expected for small  $\epsilon$ . The case  $r = 0$  (geometric mean) cannot be directly attained by the same GCLT approach, but Fisher's method provides the exact solution anyway.

Define the random variables  $X_i = p_i^r$  and

$$Y = X_1 + \dots + X_K = K M_{r,K}(p_1, \dots, p_K)^r. \quad (7)$$

Assuming independence between  $p_1, \dots, p_K$ , GCLT states<sup>21</sup> that

$$\frac{Y - a_{r,K}}{b_{r,K}}$$

converges to a Stable distribution with heavy tail index  $\lambda > 0$ , where  $\lambda \geq 2$  corresponds to the Normal distribution. The heavy tail index is determined by the tails of the individual  $X_i$ s, characterised as

$$\Pr(X_i > x) \approx cx^{-\lambda}, \quad x \rightarrow \infty \quad (8a)$$

$$\Pr(X_i < -x) \approx d|x|^{-\lambda}, \quad x \rightarrow \infty \quad (8b)$$

Assuming that  $p \sim U(0, 1)$ ,

$$\begin{aligned} \Pr(X_i > x) &= \Pr(p_i^r > x) \\ &= \begin{cases} \Pr(p_i > x^{1/r}) & \text{if } r > 0 \\ \Pr(p_i < x^{1/r}) & \text{if } r < 0 \end{cases} \\ &= \begin{cases} 1 - x^{1/r} & \text{if } r > 0 \text{ and } 0 \leq x \leq 1 \\ x^{1/r} & \text{if } r < 0 \text{ and } x \geq 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (9)$$

In other words,

$$X_i \sim \begin{cases} \text{Beta}(1/r, 1) & \text{when } r > 0 \\ \text{Pareto}(1, -1/r) & \text{when } r < 0 \end{cases} \quad (10)$$

Thus  $c = 1$ ,  $d = 0$  and

$$\lambda = \begin{cases} \infty & \text{if } r > 0 \\ -1/r & \text{if } r < 0 \end{cases} \quad (11)$$

Uchaikin and Zolotarev Table 2.1<sup>21</sup> gives coefficients for the corresponding Extremal Stable distributions that occur when  $(c - d)/(c + d) = 1$  (Table 1). In Table 1,  $S_{\lambda,1}$  is an Extremal Stable distribution in Nolan's S1 parameterization (Equation 1 of 22) with parameters  $\alpha = \lambda$ ,  $\beta = 1$ ,  $\sigma = 1$ ,  $\mu = 0$ . This is equivalent to Nolan's S0 parameterization (Equation 3 of 22 but it seemed to me there was a missing factor of  $i$  on line 2, second term in square brackets) with parameters  $\alpha = \lambda$ ,  $\beta = 1$ ,  $\sigma = 1$  and, if  $\alpha = 1$ ,  $\mu^0 = 0$ , or if  $\alpha \neq 1$ ,  $\mu^0 = \beta \tan(\pi\lambda/2)$ .

The moments required by Table 1 are

$$\mathbb{E}[X] = \begin{cases} \infty & \text{if } r \leq -1 \\ \frac{1}{1+r} & \text{if } r > -1 \end{cases} \quad (12)$$

$$\mathbb{V}[X] = \begin{cases} \infty & \text{if } r \leq -\frac{1}{2} \\ \frac{r^2}{(1+r)^2(1+2r)} & \text{if } r > -\frac{1}{2} \end{cases} \quad (13)$$

The main result of this paper is that a general significance threshold for the GMP with exponent  $r$  based on GCLT is therefore

**Table 1.** Generalized central limit theorem results for non-negative random variables<sup>21</sup>.

$r$	$\lambda$	$a_{r,K}$	$\mathbf{C}_{r,K} = \mathbf{b}_{r,K}/K^{\max\{\frac{1}{2}-r\}}$	$\frac{1}{K} \sum_{i=1}^K \mathbf{x}_i$
$r < -1$	$0 < \lambda < 1$	0	$\left[ \frac{2}{\pi} \Gamma\left(-\frac{1}{r}\right) \sin\left(-\frac{\pi}{2r}\right) \right]^r$	$C_{r,K} K^{-r-1} S_{-1/r,1}$
$r = -1$	$\lambda = 1$	$K \log K$	$\frac{\pi}{2}$	$\frac{\pi}{2} S_{1,1} + \log K$
$-1 < r < -\frac{1}{2}$	$1 < \lambda < 2$	$K \mathbb{E}[X]$	$\left[ \frac{2}{\pi} \Gamma\left(-\frac{1}{r}\right) \sin\left(-\frac{\pi}{2r}\right) \right]^r$	$C_{r,K} K^{-r-1} S_{-1/r,1} + \mathbb{E}[X]$
$r = -\frac{1}{2}$	$\lambda = 2$	$K \mathbb{E}[X]$	$(\log K)^{1/2}$	$\sqrt{\frac{\log K}{K}} S_{2,1} + \mathbb{E}[X]$
$-\frac{1}{2} < r < 0$ $r > 0$	$\lambda > 2$	$K \mathbb{E}[X]$	$(\mathbb{V}[X]/2)^{1/2}$	$\sqrt{\frac{\mathbb{V}[X]}{2K}} S_{2,1} + \mathbb{E}[X]$

$$\Psi_{\text{GCLT},r,K}(\epsilon) = \begin{cases} \left[ \frac{a_{r,K} + b_{r,K} F_{-1/r,1}^{-1}(1-\epsilon)}{K} \right]^{1/r} & \text{if } r < 0 \\ \left[ \frac{a_{r,K} + b_{r,K} F_{2,1}^{-1}(\epsilon)}{K} \right]^{1/r} & \text{if } r > 0 \end{cases} \quad (14)$$

where  $F_{\lambda,1}^{-1}$  is the inverse cumulative distribution function of an Extremal Stable random variable  $S_{\lambda,1}$  and the coefficients  $a_{r,K}$  and  $b_{r,K}$  are defined in Table 1. It is notable that for any  $\lambda > 2$ ,  $S_{\lambda,1} \stackrel{d}{=} S_{2,1}$ , which is a Normal(0, 2) distribution. This occurs when  $r > -1/2$  because the  $p$ 's are no longer heavy-tailed in the sense that their variances are defined. Those results are therefore equivalent to a straightforward application of central limit theorem. This transition in the behaviour of the GMP has implications for its robustness to dependence.

### 2.1 Small $\epsilon$ approximation

By the theory of regularly varying functions, the general significance threshold (Equation 14) simplifies when  $r < -\frac{1}{2}$  (i.e.  $\lambda < 2$ ) to

$$\Psi_{\text{GCLT},r,K}(\epsilon) \rightarrow \frac{\epsilon}{K^{1+1/r}} \quad \text{as } \epsilon \rightarrow 0 \quad (15)$$

because, (see Davis and Resnick<sup>18</sup> Lemma 2.1)

$$\begin{aligned} \Pr\left(M_{r,K}(p_1, \dots, p_K) < \frac{\epsilon}{K^{1+1/r}}\right) &= \Pr\left(K M_{r,K}(p_1, \dots, p_K)^r > K^{-r} \epsilon^r\right) \\ &= \Pr(p_1^r + \dots + p_K^r > K^{-r} \epsilon^r) \\ &\rightarrow K \Pr(X > K^{-r} \epsilon^r) \quad \text{as } \epsilon \rightarrow 0 \\ &= K(K^{-r} \epsilon^r)^{1/r} = \epsilon. \end{aligned} \quad (16)$$

subject to the Davis-Resnick condition (Equation 6).

The small  $\epsilon$  approximation shows that the HMP is the only GMP that can be directly interpreted as if it were a  $p$ -value, and only then when  $\epsilon \rightarrow 0$ . The small  $\epsilon$  approximation is compared to the significance thresholds of Vovk and Wang<sup>15</sup> in Table 2.

**Table 2.** Significance thresholds for  $M_{r,K}(p_1, \dots, p_K)$  assuming large  $K$  and, for the GCLT threshold, small  $\epsilon$ .

$r$	$\lambda$	$\Psi_{\text{RRA},r,K}(\epsilon)^{15}$	$\Psi_{\text{GCLT},r,K}(\epsilon)$ as $\epsilon \rightarrow 0$ (Equation 15)
$r < -1$	$0 < \lambda < 1$	$\frac{\epsilon}{\frac{r}{r+1} K^{1+1/r}}$	$\frac{\epsilon}{K^{1+1/r}}$
$r = -1$	$\lambda = 1$	$\frac{\epsilon}{\log K}$	$\epsilon$
$-1 < r < -\frac{1}{2}$	$1 < \lambda < 2$	$\frac{\epsilon}{(r+1)^{1/r}}$	$\frac{\epsilon}{K^{1+1/r}}$
$r \geq -\frac{1}{2}$	$\lambda \geq 2$	$\frac{\epsilon}{(r+1)^{1/r}}$	no small $\epsilon$ approx.

Equation 16 appears to be applicable in the transitional region  $-\frac{1}{2} \leq r < 0$  (i.e.  $\lambda \geq 2$ ) but here the tail behaviour can be alternatively characterised as Gaussian. Empirically, both approximations appear to struggle, so I caution that the small  $\epsilon$  approximation is not helpful for  $r \geq -\frac{1}{2}$ .

For small  $\epsilon$ , the GCLT significance threshold is less stringent than the RRA significance threshold by a factor of  $r/(1+r)$  when  $r < -1$ ,  $\log K$  when  $r = -1$  and  $(r+1)^{1/r}/K^{1+1/r}$  when  $-1 < r < -1/2$ . As  $r \rightarrow -\infty$ , the GCLT and RRA thresholds converge to those of the Šidák and Bonferroni procedures respectively, which are equivalent for small  $\epsilon$ . Even below  $r = -2$ , the difference in stringency is less than two-fold when  $\epsilon$  is small, suggesting that in this region, the approaches are similar. However, directly comparing the significance thresholds only allows a comparison of the extreme cases of independence and arbitrary dependence. More generally, an explicit model of  $p$ -value dependence is required, which is the subject of the next section.

### 3 Dependence structure of likelihood ratio tests

In motivating a dependence structure for  $p$ -values, I consider the  $p$ -values to have arisen from nested likelihood ratio tests, in which each  $p$ -value is a regularly varying function of the maximized likelihood ratio  $R_i$  for a pair of nested models  $\mathcal{M}_0$  and  $\mathcal{M}_i$ . Asymptotic theory for classical inference states (under various assumptions<sup>23</sup>) that the deviance equals

$$2\log R_i \approx \mathbf{S}_i' \mathbf{S}_i \quad (17)$$

$$= \left( \mathbf{V}[\hat{\theta}_i]^{-1/2} (\hat{\theta}_i - \theta_{0i}) \right)' \left( \mathbf{V}[\hat{\theta}_i]^{-1/2} (\hat{\theta}_i - \theta_{0i}) \right), \quad (18)$$

where  $\hat{\theta}_i$  is the maximum likelihood estimate of the  $v_i$  parameters to be estimated under model  $\mathcal{M}_i$  but not  $\mathcal{M}_0$ ,  $\theta_{0i}$  are their assumed values under  $\mathcal{M}_0$ , and  $\mathbf{V}[\hat{\theta}_i]$  is the variance-covariance matrix of the maximum likelihood estimate. Under the usual assumptions,  $\mathbf{V}[\hat{\theta}_i]$  is a function of the Fisher information matrix.

Since asymptotically,  $\mathbb{E}[\hat{\theta}_i] = \theta_i$ , the true value of the parameter, then under the null hypothesis,  $\mathbf{S}_i \sim \text{Normal}_{v_i}(\mathbf{0}, \mathbf{I})$ . I.e.  $\mathbf{S}_i$  follows an uncorrelated standard multivariate normal distribution, and  $2\log R_i$  follows a Chi-Squared( $v_i$ ) distribution.

The above outline implies that

$$\begin{pmatrix} (\hat{\theta}_1 - \theta_{01}) \\ \vdots \\ (\hat{\theta}_K - \theta_{0K}) \end{pmatrix} \sim \text{Normal}_{v_1 + \dots + v_K} \left( \begin{pmatrix} \theta_1 - \theta_{01} \\ \vdots \\ \theta_K - \theta_{0K} \end{pmatrix}, \begin{pmatrix} \mathbf{V}[\hat{\theta}_1] & \cdots & \mathbb{C}[\hat{\theta}_1, \hat{\theta}_K] \\ \vdots & \ddots & \vdots \\ \mathbb{C}[\hat{\theta}_K, \hat{\theta}_1] & \cdots & \mathbf{V}[\hat{\theta}_K] \end{pmatrix} \right), \quad (19)$$

where  $\mathbb{C}$  represents the covariance. Therefore

$$\begin{pmatrix} S_1 \\ \vdots \\ S_K \end{pmatrix} = \begin{pmatrix} \mathbb{V}[\hat{\theta}_1]^{-1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbb{V}[\hat{\theta}_K]^{-1/2} \end{pmatrix} \begin{pmatrix} (\hat{\theta}_1 - \theta_{01}) \\ \vdots \\ (\hat{\theta}_K - \theta_{0K}) \end{pmatrix} \quad (20)$$

$$\sim \text{Normal}_{v_1+\dots+v_K} \left( \begin{pmatrix} \mathbb{V}[\hat{\theta}_1]^{-1/2}(\theta_1 - \theta_{01}) \\ \vdots \\ \mathbb{V}[\hat{\theta}_K]^{-1/2}(\theta_K - \theta_{0K}) \end{pmatrix}, \begin{pmatrix} \mathbf{I} & \cdots & \text{Cor}[\hat{\theta}_1, \hat{\theta}_K] \\ \vdots & \ddots & \vdots \\ \text{Cor}[\hat{\theta}_K, \hat{\theta}_1] & \cdots & \mathbf{I} \end{pmatrix} \right) \quad (21)$$

where  $\text{Cor}[\hat{\theta}_i, \hat{\theta}_j] = \mathbb{V}[\hat{\theta}_i]^{-1/2} \mathbb{C}[\hat{\theta}_i, \hat{\theta}_j] \mathbb{V}[\hat{\theta}_j]^{-1/2}$ . Thus,  $S$  has a multivariate Normal distribution with variance-covariance matrix equal to the block matrix of individual correlation matrices between the maximum likelihood estimates of each pair of models. Call that matrix  $\rho$ . The matrix might be positive-semi-definite, rather than positive-definite, because of collinearity between the  $K$  models.

When the null hypothesis is true (i.e.  $\theta_i = \theta_{0i} \forall i = 1 \dots K$ ),

$$SS' \sim \text{Wishart}_{v_1+\dots+v_K}(\rho, 1). \quad (22)$$

The diagonal of  $SS'$ , which has a Wishart-Multivariate-Gamma distribution<sup>24</sup>, models the dependence within and between the terms in all the sums  $2 \log R_i = S_i' S_i, i = 1 \dots K$ . However, the analytical results for this distribution are limited, so in practice Equation 21 is used for simulation. After computing the maximized likelihood ratios  $R_1, \dots, R_K$  via Equation 17, the  $p$ -values are computed from the quantile functions of the corresponding Chi-Squared( $v_i$ ) distributions.

### 3.1 Simplified Wishart-Multivariate-Gamma dependence

For the simulations to test the power and false positive rate of the GMP significance thresholds, I used a simplification of the Wishart-Multivariate-Gamma dependence structure with a single parameter,  $0 \leq \rho \leq 1$ , which measures the strength of positive dependence between the log-likelihood ratios, and hence the  $p$ -values. I made the simplifying assumption that  $v_i = v$  for all  $i = 1 \dots K$  and (for  $i \neq j$ )

$$\text{Cor}[\hat{\theta}_i, \hat{\theta}_j] = \begin{pmatrix} \rho & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \rho \end{pmatrix}. \quad (23)$$

In this scenario, every alternative hypothesis has  $v$  free parameters compared to its nested null hypothesis. These might be considered to represent parameters that are in some way analogous from test to test. For any pair of likelihood ratio tests, estimates of the analogous parameters are correlated (with correlation coefficient  $\rho$ ) but estimates of non-analogous parameters are uncorrelated. In which case the joint distribution of  $(2 \log R_1, \dots, 2 \log R_K)'$  when the null hypothesis is true is modelled by the diagonal elements of

$$\text{Wishart}_K \left( \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}, v \right). \quad (24)$$

In particular, I took  $v = 2$ , which produces the simple relationship  $p_i = 1/R_i$ . While the parameter  $\rho$  can be seen as characterizing the strength of the dependence from mild to strong, all models with  $\rho > 0$  could be considered as representing a particularly 'dense' form of dependence in which every  $p$ -value is equally correlated with every other one.

## 4 Power-robustness trade-offs

### 4.1 Independence versus arbitrary dependence

To assess the relative performance of the GCLT and RRA significance thresholds for GMPs, I began by directly comparing the thresholds themselves. This allowed me to assess the false positive rates of the RRA threshold under the assumption of independence and the GCLT threshold under worst case dependence.



For a target false positive rate,  $\epsilon$ , the GCLT and RRA significance thresholds for the GMP are written  $\Psi_{\text{GCLT},r,K}(\epsilon)$  and  $\Psi_{\text{RRA},r,K}(\epsilon)$  respectively. Two quantities can be studied easily:

1. The false positive rate of the RRA threshold under independence:

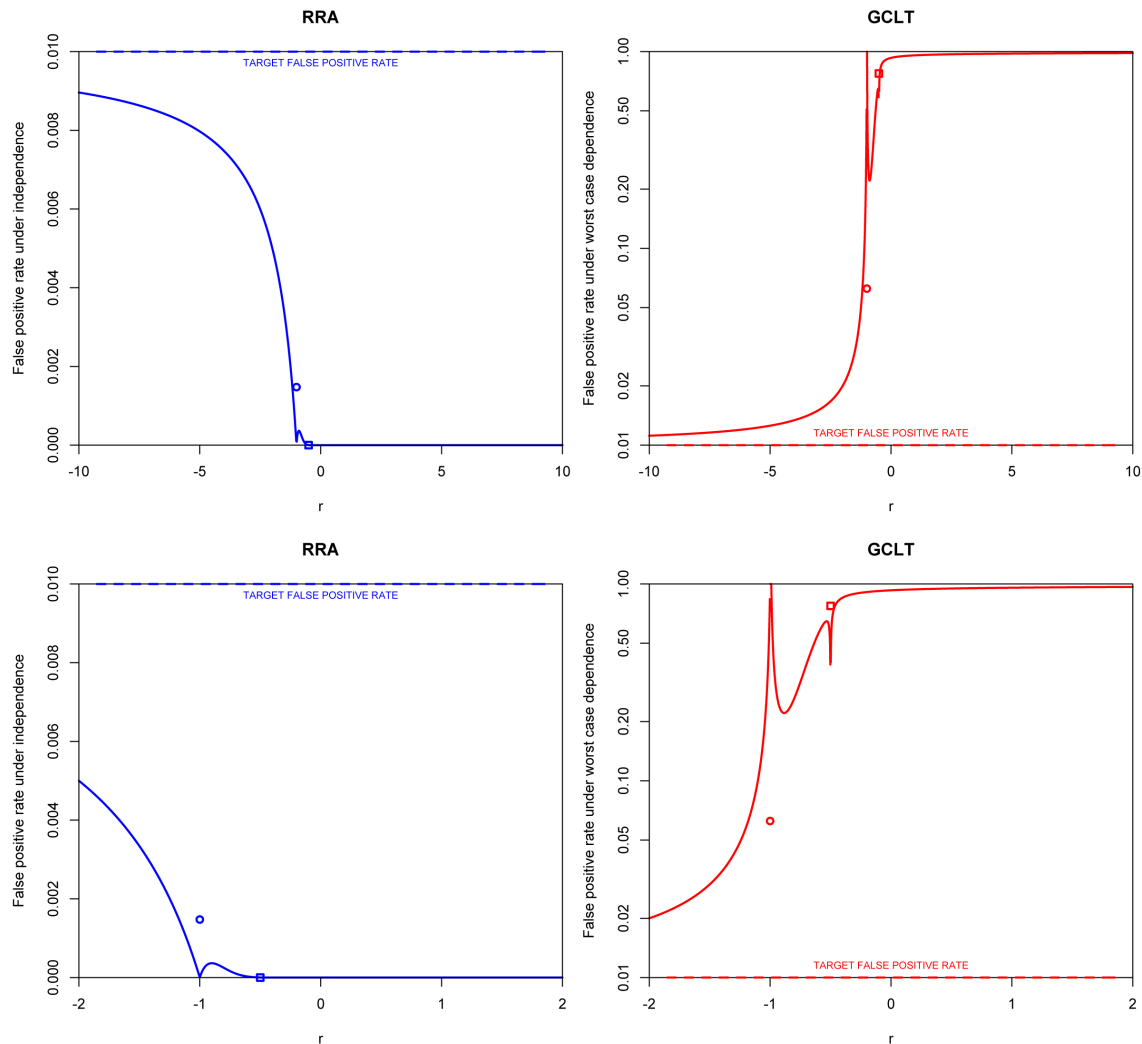
$$\Psi_{\text{GCLT},r,K}^{-1}(\Psi_{\text{RRA},r,K}(\epsilon)) \quad (25)$$

2. The false positive rate of the GCLT threshold under worst case dependence (assuming the RRA thresholds are precise):

$$\Psi_{\text{RRA},r,K}^{-1}(\Psi_{\text{GCLT},r,K}(\epsilon)) \quad (26)$$

The two quantities are expected to be below and above  $\epsilon$  respectively. Neither extreme (independence nor worst case dependence) is thought to represent empirical dependence: Wishart-Multivariate-Gamma dependence scenarios are considered later.

The two quantities are plotted in Figure 1 for  $\epsilon = 0.01$  and  $K = 1000$  over a range of  $r$ . The plots support the idea that large negative values of  $r$  produce tests that are most robust to assumptions regarding



**Figure 1. Trade-offs in robustness of GMP significance thresholds to dependence.** Two ranges are shown:  $r \in [-10, 10]$  (top panels) and  $r \in [-2, 2]$  (bottom panels). False positive rates using the RRA thresholds assuming independence (left panels, blue lines) and false positive rates using GCLT thresholds assuming worst case dependence and assuming the RRA bounds are precise (right panels, red lines). A target false positive rate of  $\epsilon = 0.01$  and  $K = 1000$  tests were assumed. Discontinuities occur at  $r = -1$  (the HMP) and  $r = -\frac{1}{2}$ . The false positive rates at these positions are marked by a circle and square respectively. Note the logarithmic y-axis on the right panels. R code is available as *Extended data* on Figshare: [doi:10.6084/m9.figshare.11907033](https://doi.org/10.6084/m9.figshare.11907033)<sup>25</sup>.

dependence and significance thresholds that are most similar. In the extreme case that  $r \rightarrow -\infty$ , this implies use of Bonferroni and Šidák correction under arbitrary dependence and independence respectively: the resulting significance thresholds converge for small  $\epsilon$ .

**Figure 1** also visualises the transition that occurs at  $r = -\frac{1}{2}$  between the heavy-tailed ( $r < -\frac{1}{2}$ ) and light-tailed ( $r \geq -\frac{1}{2}$ ) distributions of  $p^r$ . Below  $r = -\frac{1}{2}$  the heavy tails of the individual  $p^r$ s result in convergence of  $M_{r,K}(p_1, \dots, p_K)^r$  to a Stable distribution, whereas above  $r = -\frac{1}{2}$  it converges to a Normal distribution. The Davis-Resnick condition implies that that sum of heavy-tailed random variables is robust to dependence in the tail: **Figure 1** (right side) shows that robustness to arbitrary dependence is a process that begins at  $r = -\frac{1}{2}$  but requires  $r$  substantially below  $-\frac{1}{2}$  to become appreciable. For example, under worst case dependence, the false positive rate of the HMP ( $r = -1$ ) is still elevated 6.2-fold above the target false positive rate,  $\epsilon = 0.01$ .

As the exponent falls to  $r = -2$ , the inflation in false positive rate of the GCLT threshold above its target drops to two-fold, a notable value that corresponds to the worst case for *direct interpretation* of the arithmetic mean  $p$ -value (AMP,  $r = 1$ )<sup>26,27</sup>. If, however, one applies the GCLT significance threshold to the AMP, rather than directly interpreting it, the false positive rate jumps to 0.96 under worst case dependence because direct interpretation of the AMP is highly conservative under independence. The disparity in false positive rates illustrates the vastly superior robustness to dependence of the GCLT threshold, and the GMP in general, at  $r = -2$  versus  $r = 1$ . It also shows that direct interpretation the AMP is questionable: not only is it up to two-fold anti-conservative under worst case dependence between tests, but its power will be highly compromised for independent tests. GMPs with smaller exponents have intrinsically greater robustness to dependence.

Robustness to dependence is a desirable property in the false positive rate, but there may be a cost in terms of the power to reject the null hypothesis when it is false. The well-known conservatism of Bonferroni correction suggests this is inevitable.

## 4.2 Simulations under Wishart-Multivariate-Gamma dependence

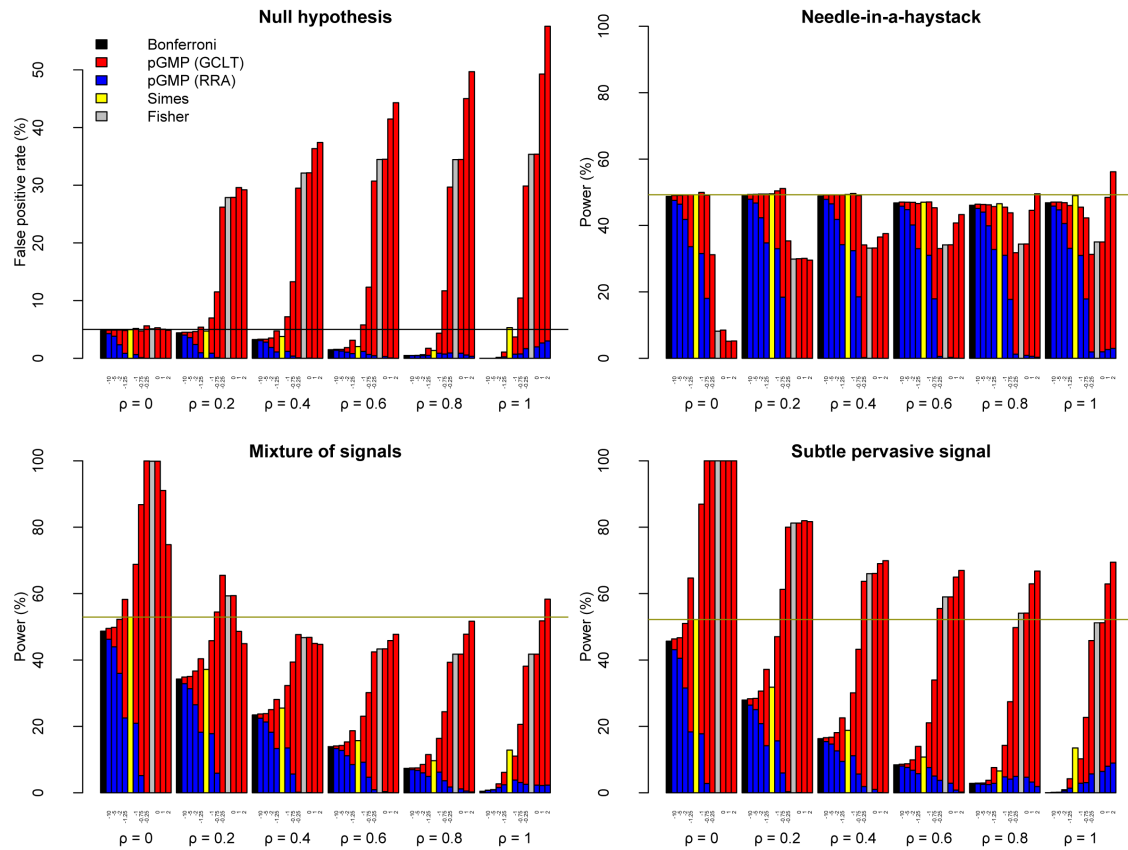
For a representative evaluation of test performance it is necessary to consider empirically relevant dependence and to compare not just false positive rates, but power. In this section, I report simulations that I conducted under Wishart-Multivariate-Gamma dependence, a form of dependence motivated earlier by the asymptotic distribution of log-likelihood ratios among dependent tests. I considered a simplified form of Wishart-Multivariate-Gamma dependence in which all tests were equally correlated, with a single correlation parameter  $\rho$  (defined in [section 3.1](#)).

To evaluate false positive rate and power, I considered four scenarios: **Null hypothesis**, **Needle-in-a-haystack**, **Mixture of signals** and **Subtle pervasive signal**. The scenarios differed in the number of  $p$ -values, out of 1000, simulated under the alternative hypothesis (0, 1, 100 and 1000 respectively), and in the value of the Z-statistic ( $\mathbb{V}[\hat{\theta}_i]^{-1/2}(\theta_i - \theta_{0i})$ ; see [section 3](#)) when it was non-zero ( $n/a$ , 3.0, 1.25 and 0.7 respectively). The exact values were arbitrary and chosen to produce power  $\sim 50\%$  under independence. Beside the GMP tests, I conducted Bonferroni, Simes and Fisher tests for comparison. In all cases, I assumed a target false positive rate of  $\epsilon = 0.05$ .

**False positive rates.** Using the GCLT thresholds, the GMP with  $r < -1$  exhibited false positive rates that were close to the target of  $\epsilon = 5\%$  under mild dependence ( $\rho \leq 0.2$ ) and substantially below it under strong dependence ( $\rho \geq 0.6$ ) (**Figure 2**). The GMP with  $r > -1$  exhibited substantial inflation of false positive rates under mild to strong dependence. In contrast, the HMP ( $r = -1$ ) maintained a false positive rate close to the target in all cases, erring on the side of inflation for  $0.2 \leq \rho \leq 0.6$ , as observed previously<sup>17,19</sup>.

Using the RRA thresholds, the false positive rate of the GMP was always well-controlled below the target of  $\epsilon = 5\%$ , as expected, and therefore below the rates achieved using the GCLT thresholds. Often it was far below the target false positive rate. Usually it was even below the false positive rate of the Bonferroni procedure (**Figure 2**).

**Power.** In considering power, the GCLT thresholds for  $r > -1$  can be disregarded as inadmissible when  $\rho \neq 0$  because they could not control the false positive rate close to the target. In this range, the RRA thresholds were admissible but exhibited very poor power under all scenarios (**Figure 2**). The conclusion that neither the GCLT nor RRA thresholds are useful in the range  $-1 < r < -1/2$  is relevant to the later section on the interpretation of the GMP ([section 6](#)). From a practical perspective, these simulations suggest that GMPs with  $r > -1$ , including the geometric mean  $p$ -value ( $r = 0$ ; Fisher's method) and the arithmetic mean  $p$ -value ( $r = 1$ ), are not useful unless independence or an explicit model of dependence can be safely assumed, e.g. [7,16](#).

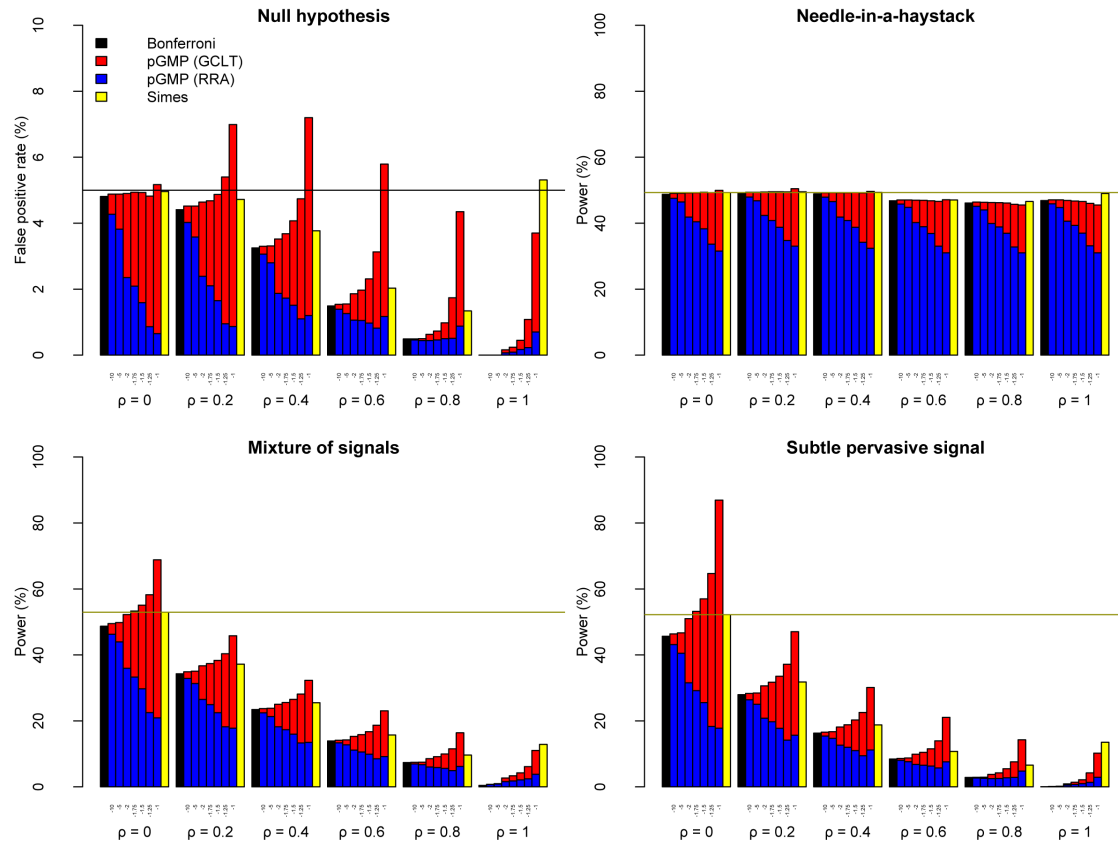


**Figure 2. Power-robustness trade-offs for the GMP and related tests.** For  $r = \{-10, -5, -2, -1.25, -1, -0.75, -0.25, 10^{-6}, 1, 2\}$  and  $\rho = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ , I conducted 10000 simulations of  $K = 1000$   $p$ -values under four scenarios (see text). For each scenario, I computed the proportion of simulations in which the GMP was below the significance threshold, calculated by GCLT (red) or RRA (blue). For comparison, I computed the proportion of simulations in which the Bonferroni (black), Simes (yellow) and Fisher (grey) tests were significant. These were plotted next to the GMP they most closely resemble:  $r = -10, -1$  and  $10^{-6}$  respectively. In all cases, I assumed a target false positive rate of  $\epsilon = 5\%$  (horizontal black line). For comparison, the dark yellow horizontal line shows the power of the Simes test. The R code is available as *Extended data* from Figshare: [doi:10.6084/m9.figshare.11907033](https://doi.org/10.6084/m9.figshare.11907033)<sup>25</sup>.

In the Needle-in-a-haystack scenario, the GCLT thresholds for  $r \leq -1$  achieved power comparable to, or slightly worse than, the Bonferroni procedure. For  $-10 \leq r \leq -1$ , the RRA thresholds were worse than the Bonferroni procedure, considerably so for  $-2 \leq r \leq -1$  (Figure 3).

In both the Mixture of signals and Subtle pervasive signal scenarios, the power of the GCLT thresholds increased from  $r = -10$  to  $r = -1$ . The trend was reversed for the RRA thresholds (except for  $\rho = 1$ ). The power of the RRA thresholds was uniformly worse (often substantially worse) than for the GCLT thresholds, and it was usually worse than for the Bonferroni thresholds too. The effect of increasing the strength of dependence was to reduce power for the GCLT, RRA and Bonferroni thresholds.

The Simes test resembles the HMP using the GCLT threshold, both in terms of interpretation and performance<sup>9</sup>. Simes' test had the advantage of avoiding inflation in false positive rate under the Null hypothesis with  $0.2 \leq \rho \leq 0.6$ . The power of Simes' test was no better than the HMP (except at  $\rho = 1$ ), but often it was appreciably worse in the Mixture of signals and Subtle pervasive signal scenarios. For smaller values of  $r$ , e.g.  $r = -1.25$ , the inflation in false positive rate under mild dependence was reduced compared to the HMP. The GCLT threshold at  $r = -1.25$  also generally outperformed Simes' test in terms of power, with the exception of  $\rho = 1$ . No value of  $r$  enabled the RRA thresholds to outperform the power of Simes' test.



**Figure 3. Power-robustness trade-offs for the GMP and related tests.** As for Figure 2 but considering GMP with  $r = \{-10, -5, -2, -1.75, -1.5, -1.25, -1\}$ , i.e. only very heavy-tailed distributions. R code is available as *Extended data* on Figshare: [doi:10.6084/m9.figshare.11907033](https://doi.org/10.6084/m9.figshare.11907033)<sup>25</sup>.

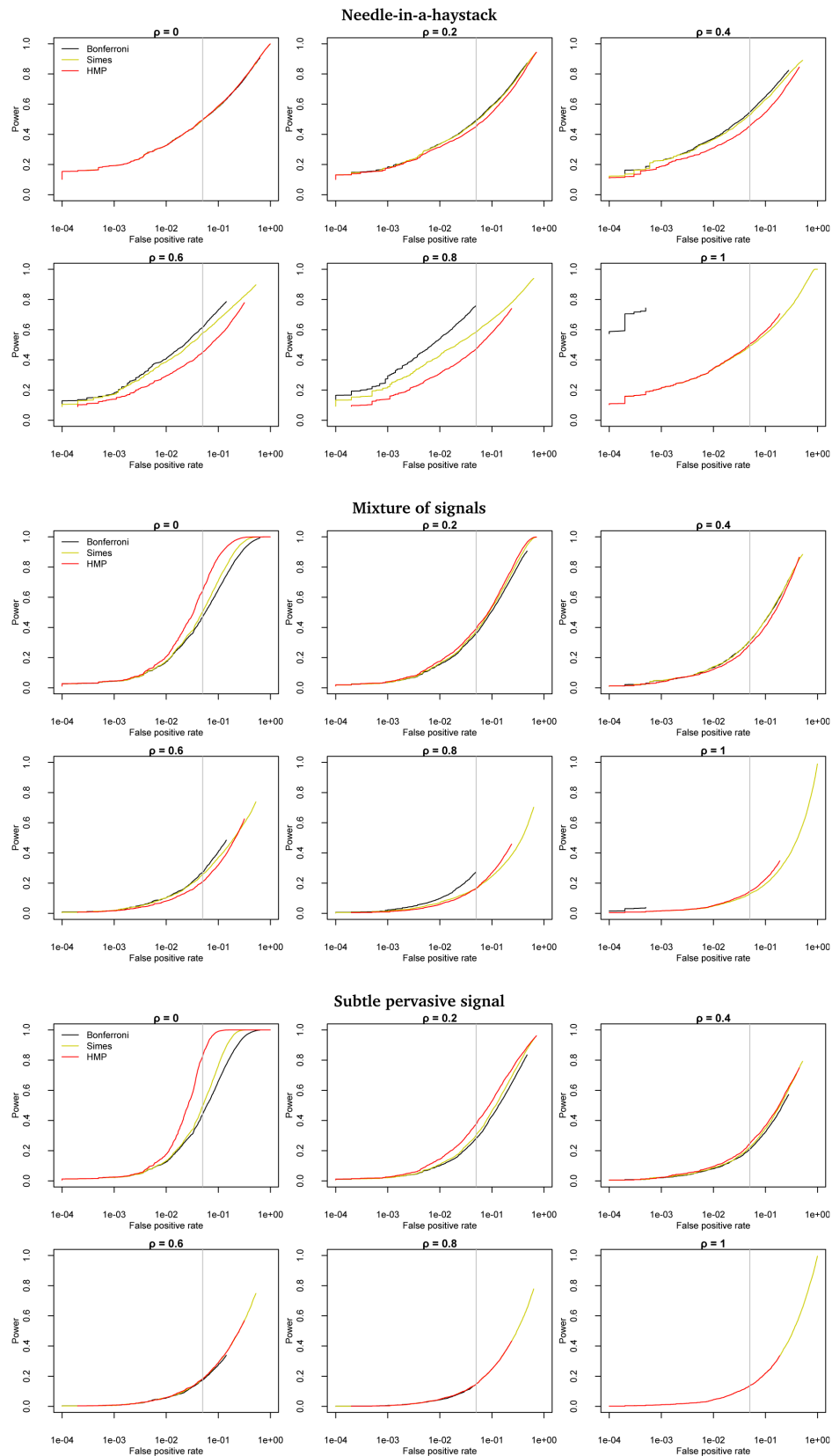
As expected, Bonferroni correction behaved much like the GMP with  $r = -10$  under both the GCLT and RRA thresholds. As expected, Fisher's method was indistinguishable from the GMP with  $r = 10^{-6}$  under the GCLT threshold. For  $\rho = 0$ , Fisher's method could not be bettered in the Mixture of signals and Subtle pervasive signal scenarios, but it was roundly outperformed in the Needle-in-a-haystack scenario (Figure 2).

In conclusion, the HMP (with GCLT threshold) and Simes' test appear to offer superior performance to the alternatives over the range of dependence structures considered. The HMP enjoys greater power than Simes' test, at the cost of an inflated false positive rate. For end-users, the relative importance of power versus a conservative false positive rate will influence the choice of test. The simulations support the claim that the HMP is robust to dependence in the sense that the realized false positive rate is close to the target across all dependence structures investigated.

#### 4.3 Inherent power of the HMP vs Simes' test

The superior power of the HMP compared to Simes' test in the presence of dependence is attributable, in part, to its higher false positive rate. However, receiver-operator curves (ROCs, Figure 4) summarizing the simulations above show that in the Mixture of signals and Subtle pervasive signal scenarios, the HMP is inherently more powerful than Simes' test. However, the advantage is reduced, and even reversed, as dependence increases. Simes' test is inherently more powerful than the HMP in the Needle-in-a-haystack scenario.

Inherent power (the power of a test when its threshold achieves the target false positive rate) cannot be realized without analytical results, which are not available for Wishart-Multivariate-Gamma dependence, or simulations. Nevertheless, one can use simulations to compare inherent power to actual power to



**Figure 4. Receiver-operator curves (power vs false positive rate) by strength of dependence and scenario.** R code is available as *Extended data* on Figshare: [doi:10.6084/m9.figshare.11907033](https://doi.org/10.6084/m9.figshare.11907033)<sup>25</sup>.

quantify the shortfall or excess power attributable to conservatism or anti-conservatism of the false positive rate. In Figure 5, the bars show inherent power while the grey whiskers compare that to actual power, using GCLT thresholds. When the grey whisker exceeded the bar, as it often did for the HMP ( $r = -1$ ), the actual power was elevated relative to inherent power because of an inflated false positive rate. When the grey whisker fell below the bar, as it usually did for Simes' test, the actual power was less than the inherent power because of a conservative false positive rate.

For  $r < -1$ , the tendency for actual power to exceed inherent power was reduced, and often reversed compared to the HMP. However, the cost of this greater robustness to dependence was reduced actual power in the Mixture of signals and Subtle pervasive signal scenarios. Empirically, the trend appeared to be monotonic except for when  $\log K/(1 - \log K) < r < -1$ , in which region the inherent and actual power were drastically reduced (not shown). Therefore,  $r = \log K/(1 - \log K)$ , which equalled  $-1.17$  for  $K = 1000$ , was the GMP with  $r < -1$  whose characteristics most closely resembled the HMP. This coincided with the derivation of the RRA threshold for  $r = -1$ , which was taken as the tightest bound based on  $r < -1$ , which occurs at  $r = \log K/(1 - \log K)$ <sup>15</sup>. Whether this recommends the use of the GMP at  $r = \log K/(1 - \log K)$  over the HMP is unclear. Figure 3 showed that GMPs with  $r < -1$  can still be subject to mildly inflated false positive rates using GCLT thresholds, even if they are attenuated relative to the HMP. Bonferroni usually over-powered the GMP using RRA thresholds.

## 5 Strong-sense family-wise error rates

One of the advantages of the HMP procedure is its ability to test arbitrary combinations of the  $K$   $p$ -values while controlling the strong-sense familywise error rate at a pre-specified level  $\epsilon$ , known as multilevel testing<sup>9</sup>. A full assessment of the relative performance of other GMPs to the HMP therefore involves a comparison of their performance as multilevel tests. This requires a closed testing procedure (CTP<sup>14</sup>) to be derived for the GMP with any exponent  $r$ .

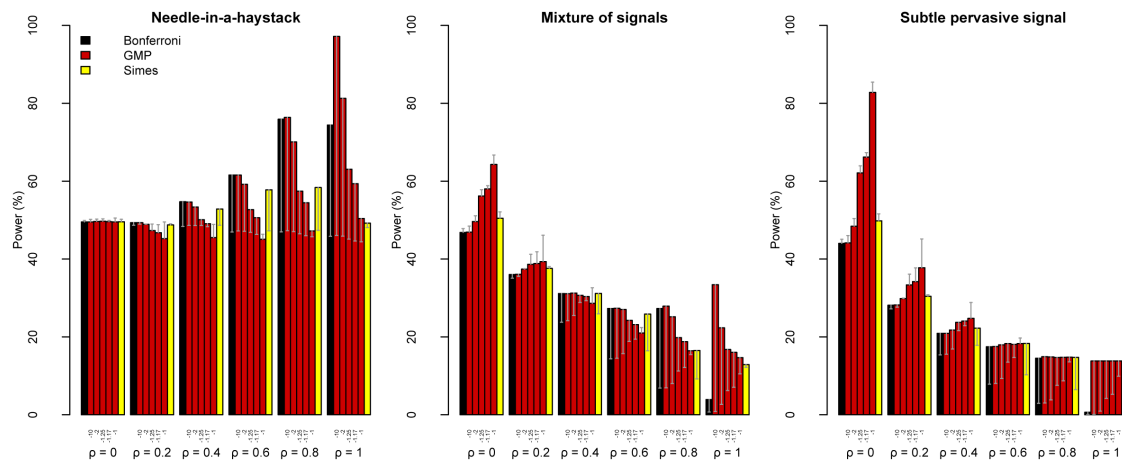
Suppose that  $\mathcal{R}$  is an index set of the  $i = 1 \dots K$   $p$ -values. As shorthand, write

$$\bar{p}_{r,\mathcal{R}} = M_{r,|\mathcal{R}|}(\{p_i : i \in \mathcal{R}\}) \quad (27a)$$

$$\bar{p}_r = M_{r,K}(p_1, \dots, p_K). \quad (27b)$$

To define the closed testing procedure of the multilevel test, find the least stringent (i.e. largest) value below 1 of the factor  $f_{|\mathcal{R}|}$  for which the following condition

$$\bar{p}_{r,\mathcal{R}} \leq f_{|\mathcal{R}|} \Psi_{r,|\mathcal{R}|}(\epsilon) \quad (28)$$



**Figure 5. Inherent versus actual power for the Bonferroni, GMP and Simes tests.** GCLT thresholds were used for the GMPs, with  $r = \{-10, -2, -1.25, -1.17, -1\}$ . Actual power (grey whiskers) exceeded the inherent power (bars) when the false positive rate was inflated relative to its target of  $\epsilon = 5\%$ . R code is available on Figshare: [doi:10.6084/m9.figshare.11907033](https://doi.org/10.6084/m9.figshare.11907033)<sup>25</sup>.

(interpreted as significance of subset  $\mathcal{R}$ ) implies the significance of all tests combined, i.e.

$$\bar{p}_r \leq \Psi_{r,K}(\epsilon). \quad (29)$$

Since

$$\bar{p}_r^r = \frac{|\mathcal{R}|}{K} \bar{p}_{r,\mathcal{R}}^r + \frac{|\mathcal{R}'|}{K} \bar{p}_{r,\mathcal{R}'}^r \quad (30)$$

then Equation 28 implies that

$$\bar{p}_r \leq \left( \frac{|\mathcal{R}|}{K} f_{|\mathcal{R}|}^r \Psi_{r,|\mathcal{R}|}(\epsilon)^r + \left( 1 - \frac{|\mathcal{R}|}{K} \right) \right)^{1/r} \quad (31)$$

assuming the least favourable case that  $\bar{p}_{r,\mathcal{R}'} = 1$ . The condition in Equation 29 is therefore satisfied by

$$f_{|\mathcal{R}|} = \min \left\{ 1, \left( \frac{\Psi_{r,K}(\epsilon)^r - \left( 1 - \frac{|\mathcal{R}|}{K} \right)}{\frac{|\mathcal{R}|}{K} \Psi_{r,|\mathcal{R}|}(\epsilon)^r} \right)^{1/r} \right\}. \quad (32)$$

The above reveals that the multilevel test suffers complications when  $r > 0$ . This is because for subsets of  $p$ -values smaller than  $K(1 - \Psi_{r,K}(\epsilon)^r)$ , the numerator of Equation 32 can be negative when  $r > 0$ . The interpretation is that there is no value of  $\bar{p}_{r,\mathcal{R}}$  small enough to compensate for the assumption that  $\bar{p}_{r,\mathcal{R}'} = 1$ . Thus no individual  $p$ -value can be sufficiently significant to guarantee that the combined test is significant. (Although knowledge of the rank of the  $p$ -value would alter this conclusion, e.g. knowing it was the maximum. The problem arises because this multilevel test is a shortcut to the full CTP and is based on single  $p$ -values. CTPs are not ruled out for  $r > 0$  in general.) From a practical perspective, it means that when  $r > 0$ , no subsets of  $p$ -values smaller than  $K(1 - \Psi_{r,K}(\epsilon)^r)$  can be significant within this multilevel test, limiting the finest levels at which inference can be made.

The multilevel test simplifies when  $r < -1$  because the GCLT thresholds (Equation 14 and Table 1) and RRA thresholds (Table 2) possess the property that  $\Psi_{r,K}(\epsilon)/\Psi_{r,|\mathcal{R}|}(\epsilon) = (|\mathcal{R}|/K)^{1+1/r}$ . This allows a more stringent form of Equation 32 to be expressed as

$$\begin{aligned} f_{|\mathcal{R}|} &= \min \left\{ 1, \left( \frac{\Psi_{r,K}(\epsilon)^r}{\frac{|\mathcal{R}|}{K} \Psi_{r,|\mathcal{R}|}(\epsilon)^r} \right)^{1/r} \right\}, \quad r < 0 \\ &= \frac{|\mathcal{R}|}{K}, \quad r < -1. \end{aligned} \quad (33)$$

This produces a convenient form of the CTP:

$$\bar{p}_{r,\mathcal{R}} \leq \left( \frac{|\mathcal{R}|}{K} \right) \Psi_{r,|\mathcal{R}|}(\epsilon) = \left( \frac{|\mathcal{R}|}{K} \right)^{-1/r} \Psi_{r,K}(\epsilon), \quad r < -1. \quad (34)$$

Thus by RRA one has,

$$\bar{p}_{r,\mathcal{R}} \leq \frac{|\mathcal{R}|^{-1/r}}{K} \frac{r+1}{r} \epsilon, \quad r < -1 \quad (35)$$

and by GCLT (Equation 16 and 9), one has a simple small  $\epsilon$  approximation

$$\bar{p}_{r,\mathcal{R}} \leq \frac{|\mathcal{R}|^{-1/r}}{K} \epsilon, \quad r \leq -1, \quad \epsilon \rightarrow 0. \quad (36)$$

Further, Equation 34 shows that all CTPs in the range  $-\infty < r < -1$  have a cost relative to Bonferroni ( $r \rightarrow -\infty$ ) in the sense that the significance threshold for an individual  $p$ -value is more stringent by a factor  $\frac{r+1}{r}$  (by RRA) or  $\epsilon^{-1} (C_{-1/r} F_{-1/r,1}^{-1}(\epsilon))^{1/r}$  (by GCLT), although Equation 36 shows that the latter is close to 1 for small  $\epsilon$ . Intuitively, this represents the cost of the additional power to make statements about *groups* of  $p$ -values over and above the statements one can make about *individual*  $p$ -values<sup>17</sup>. However, the multilevel Simes test does not incur this penalty.

## 6 Interpretation

The  $p$ -value can be seen as a low-dimensional summary of the data that is relevant to hypothesis testing. From this perspective, the distribution of the  $p$ -value under the alternative can be modelled directly, e.g. 28. Heard and Rubin-Delanchy<sup>4</sup> considered Beta distributions for the  $p$ -value under the alternative,

$$p_i | \mathcal{M}_i \sim \text{Beta}(\xi, \zeta),$$

subject to the constraint that the density is non-increasing in  $p$ , which implies that  $0 < \xi \leq 1$  and  $1 \leq \zeta$ . By the Neyman-Pearson lemma<sup>29</sup>, they argued that one can identify uniformly most powerful tests for combining independent  $p$ -values under the Beta distribution assumption. Fisher's method was optimal for  $\zeta = 1$ , the subset of Beta distributions that have been advocated for local alternatives<sup>30</sup>. Pearson's method<sup>31</sup> was optimal for  $\xi = 1$ , the subset of Beta distributions that have been advocated for simple alternatives<sup>30</sup>.

The likelihood ratio for a  $p$ -value that is Beta distributed under the alternative can be written

$$\text{BF}_i = \frac{p_i^{\xi-1} (1-p_i)^{\zeta-1}}{\text{B}(\xi, \zeta)}, \quad (37)$$

where  $\text{B}(\cdot)$  is the Beta function. The notation  $\text{BF}_i$  is used because the local alternatives assumption amounts to a Bayesian prior distribution over effect sizes with hyper-parameter  $\xi$ , and the likelihood ratio is therefore a Bayes factor.

A mean Bayes factor (or likelihood ratio) of the form

$$\overline{\text{BF}} = \frac{1}{K} \sum_{i=1}^K \frac{p_i^{\xi-1} (1-p_i)^{\zeta-1}}{\text{B}(\xi, \zeta)} \quad (38)$$

arises in a model-averaging setting in which the alternative hypothesis is a mixture of individual, mutually exclusive alternatives. Here each  $p$ -value uses the *same* data to evaluate each competing alternative hypotheses against a common nested null hypothesis. This implies an interpretation of GMPs as uniformly most powerful tests for model-averaged alternative hypotheses, each of which is a different local alternative to the common nested null hypothesis. Under these conditions,  $\xi = 1 + r$  and  $\zeta = 1$  so that

$$M_{r,K}(p_1, \dots, p_K)^r = \overline{\text{BF}} = \frac{1}{K} \sum_{i=1}^K (1+r) p_i^r. \quad (39)$$

The model-averaging interpretation applies when  $-1 < r < 0$ . Unfortunately, the simulations summarized in Figure 2 showed that for  $-1 < r < 0$ , the GCLT threshold suffered greater elevation in false positive rates than



the HMP. (The RRA threshold is not considered here because power was low and usually worse than Bonferroni). Nevertheless, outside the range  $-1 < r < 0$ , the GMP can be viewed as a bound on the model-averaged Bayes factor. Defining  $r^* < r < 0$ , one has the relationship

$$\overline{\text{BF}} \leq (1 + r) M_{r^*, K}(p_1, \dots, p_K) \quad (40)$$

because  $M_{r, K}(p_1, \dots, p_K) \geq M_{r^*, K}(p_1, \dots, p_K)$ . Therefore GMPs supply upper bounds on Bayes factors when  $r \leq -1$  and lower bounds on Bayes factors when  $r > 0$ . The HMP supplies the tightest upper bound for well-powered tests ( $r \downarrow -1$ ). Simes' test and Bonferroni can be framed similarly as providing lower bounds on the model-averaged Bayes factor.

While the HMP has a natural interpretation as a tight bound on the model-averaged Bayes factor for well-powered tests, it does not correspond exactly to any Bayes factor and it does not control the frequentist false positive rate exactly for some dependence structures, making it an approximation to both approaches that could be criticized for satisfying neither.

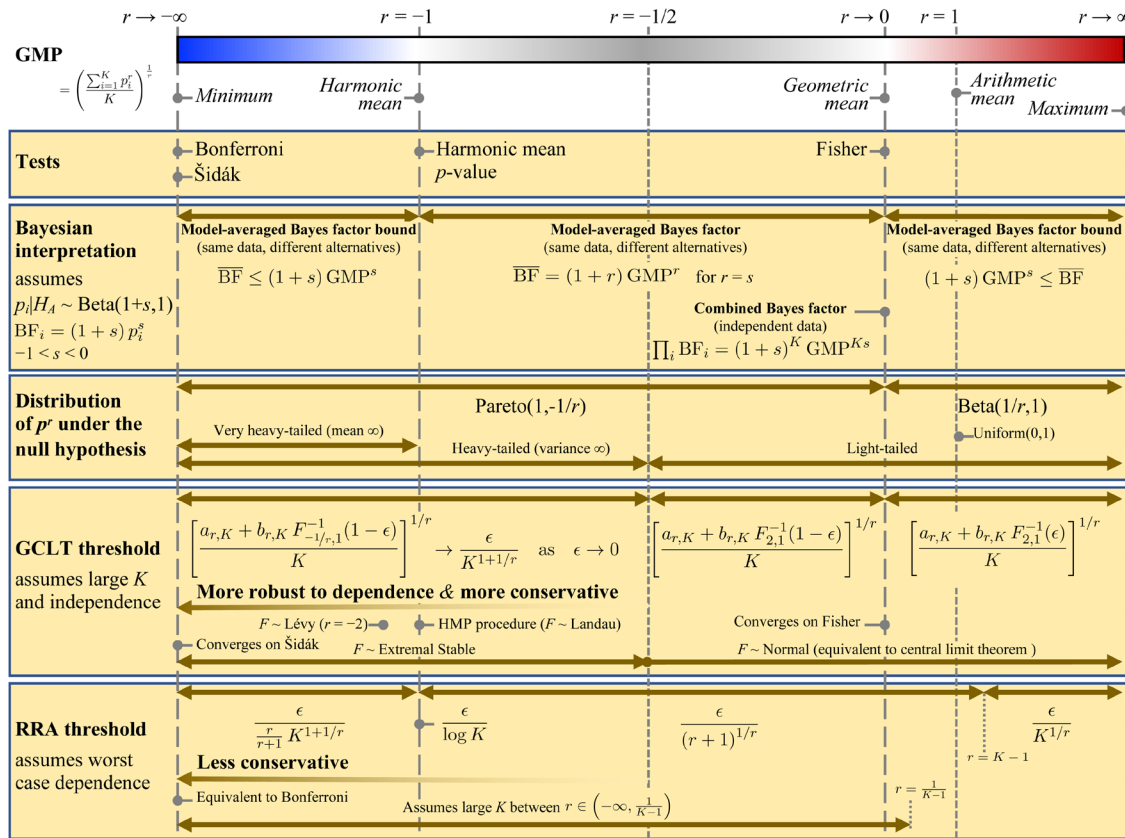
## 7 Conclusions

Taking the generalized mean  $p$ -value of a group of tests extends a number of existing methods for combining  $p$ -values including the Bonferroni, Šidák, harmonic mean  $p$ -value and Fisher procedures<sup>7,9-12</sup> (Figure 6). The interpretation varies by (i) the exponent of the GMP, and (ii) the key assumption regarding dependence between the tests. Two appealing interpretations occur when  $-1 < r < 0$  and  $r = 0$ . When  $-1 < r < 0$ , combining  $p$ -values using the GMP can be interpreted as model averaging if the same data have been used to evaluate mutually exclusive alternative hypotheses against a common null hypothesis. In this interpretation, when  $r$  is closer to  $-1$ , very small  $p$ -values are assumed more likely when the alternative hypothesis is true, implying the individual tests were more powerful. Outside the range  $-1 < r < 0$ , the GMP is interpretable as approximating this approach, with  $r = -1$  (the HMP) offering the closest approximation for well-powered tests. When  $r = 0$ , combining  $p$ -values using the GMP can be interpreted as aggregating evidence for related pairs of alternative and null hypotheses, if independent data were used for the individual tests, in which case the method is equivalent to Fisher's for many tests. Outside these specific interpretations, the GMP offers a flexible non-parametric approach to combining  $p$ -values where  $r$  controls the sensitivity to small values.

GMPs are not directly interpretable as  $p$ -values in general<sup>15</sup>. Instead, significance thresholds are required. Generalized central limit theorem and robust risk analysis provide convenient methods for defining significance thresholds for GMPs that do not require explicit knowledge of the dependence structure, providing robustness to dependence to varying degrees. RRA thresholds provide robustness to arbitrary dependence in the sense that the false positive rate will not exceed the target<sup>15</sup>. GCLT thresholds provide a weaker form of robustness to forms of dependence that satisfy the Davis-Resnick condition (Equation 6), but only for sufficiently small values of the target false positive rate and  $r \leq -1$ . Subject to these conditions, the HMP is the only GMP that can be directly interpreted as if it were a  $p$ -value<sup>9,17</sup>.

The cost of robustness to arbitrary dependence was too high to make the RRA thresholds directly useful in practice, because they were usually rendered less powerful than the Bonferroni procedure in simulations (Figure 2, Figure 3). However, they remain theoretically valuable because they bound the worst-case inflation of the false positive rate of the GCLT thresholds. The RRA and GCLT thresholds agreed more closely as  $r \rightarrow -\infty$ . The trend for RRA thresholds to deliver less powerful tests as  $r$  increased was reversed for GCLT thresholds. In practice the GCLT thresholds were generally more powerful than Bonferroni, and increasingly so as  $r$  increased, but they began to suffer inflated false positive rates. The GCLT threshold for the HMP has previously been shown to suffer modest inflation under mild dependence<sup>19</sup>. However, above  $r = -1$ , the point at which the underlying distribution of  $p_r$  transitioned from very heavy tailed ( $r \leq -1$ ) to heavy tailed ( $-1 < r \leq -1/2$ ), inflation accelerated to the point that there was no useful robustness to non-independence (Figure 2). Despite this problem, incorporating knowledge of dependence into standard central limit theorem, applicable for light-tailed distributions ( $r > -1/2$ ), would be straightforward, requiring knowledge only of  $C(p_i^r, p_j^r)$ . The loss of robustness to dependence recommends against the use of the GMP with GCLT thresholds for  $r > -1$  except when independence can be safely assumed.

The arithmetic mean  $p$ -value, which arises in numerous applications including posterior predictive  $p$ -values, is known to be directly interpretable subject to a maximum two-fold inflation in false positive rate<sup>25,27</sup>. However, for the dependence structures considered here, there appeared to be little merit in *direct interpretation* of the AMP: to do so would be far too conservative under independence and likely less powerful than Bonferroni under



**Figure 6.** Summary of the generalized mean  $p$ -value (GMP), its relation to other tests, Bayesian interpretation, significance thresholds derived using generalized central limit theorem (GCLT) and robust risk analysis (RRA)<sup>15</sup>, test assumptions and performance characteristics.

worst-case dependence. An interesting alternative might be the GMP with  $r = -2$ , whose GCLT threshold is  $\epsilon/\sqrt{K}$  for small  $\epsilon$  and does not suffer at worst two-fold inflation under arbitrary dependence (for large  $K$ ). Unlike the HMP, the GMP with  $r = -2$  did not suffer even mild inflation in false positive rate in simulations, but the HMP was more powerful (Figure 3). The GMP with  $r = -2$  performed remarkably similarly in false positive rate and power to Simes' test<sup>20</sup>. Simes' test and the HMP can be seen as offering similarly-performing but complementary solutions to the power-robustness trade-off for model-averaged  $p$ -values<sup>9</sup>, erring on the side of conservatism versus power respectively.

There were several limitations in the current study: (i) Equal weights were assumed throughout, although simulations for the HMP<sup>9</sup> suggest there may be robustness to unequal weights, at least for  $r \leq -1$ . (ii) The distribution of  $p$ -values under the null hypothesis was assumed to be Uniform(0,1). However, valid  $p$ -values are generally defined such that  $\Pr(p < x | \mathcal{M}_0) \leq x$ . Conservatism of this sort was not explored, but is likely to profoundly diminish the power of the GMP. (iii) The simulations considered here assumed a particular form of dependence in which the  $p$ -values were chi-squared tail probabilities of underlying log-likelihood ratios that for large samples would follow a Wishart-Multivariate-Gamma distribution. A particularly dense form of dependence was assumed that applied to every pair of  $p$ -values. Some results, such as the inflation in false positive rates for the GCLT thresholds in the region  $r = -1$ , will depend quantitatively on the details of the simulations. The conclusion that the RRA thresholds are less powerful than Bonferroni may apply more widely because it stems from the theoretical divergence in GCLT and RRA thresholds as  $r$  increases, and it might seem reasonable to assume that the behaviour of empirically relevant  $p$ -value dependence is intermediate between their respective assumptions of independence and arbitrary dependence.

In conclusion, simulations under a form of dependence relevant to  $p$ -values calculated from likelihood ratio tests showed that the GMP is practically useful for combining dependent  $p$ -values for exponents  $r \leq -1$  using

thresholds derived from generalized central limit theorem. Robust risk analysis provides corresponding upper bounds on the false positive rate under worst case dependence<sup>15</sup>, but these upper bounds were not directly useful as significance thresholds because they produced tests typically less powerful than Bonferroni. Those wishing to protect themselves against worst case dependence should therefore prefer the Bonferroni procedure. However, there is increasing interest in exploiting heavy tail behaviour to confer desirable properties in terms of power and robustness to dependence upon combined tests<sup>9,32</sup>, and the GMP for  $r \leq -1$  with GCLT thresholds extends this class of methods.

## 8 Data availability

### Underlying data

No underlying data are associated with this article.

### Extended data

Figshare: R code for Figures. <https://doi.org/10.6084/m9.figshare.11907033.v1><sup>25</sup>.

This project contains R code for Figure 1–Figure 5.

Extended data are available under the terms of the [Creative Commons Attribution 4.0 International license\(CC-BY 4.0\)](#).

## References

- Hedges LV, Olkin I: **Statistical methods for meta-analysis**. Academic Press, 1985.  
[Reference Source](#)
- Zaykin DV: **Optimally weighted z-test is a powerful method for combining probabilities in meta-analysis**. *J Evol Biol.* 2011; **24**(8): 1836–1841.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Loughin TM: **A systematic comparison of methods for combining p-values from independent tests**. *Comput Stat Data Anal.* 2004; **47**(3): 467–485.  
[Publisher Full Text](#)
- Heard NA, Rubin-Delanchy P: **Choosing between methods of combining p-values**. *Biometrika.* 2018; **105**(1): 239–246.  
[Publisher Full Text](#)
- Zheng J, Erzurumluoglu AM, Elsworth BL, et al.: **LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis**. *Bioinformatics.* 2017; **33**(2): 272–279.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Willer CJ, Li Y, Abecasis GR: **Metal: fast and efficient meta-analysis of genomewide association scans**. *Bioinformatics.* 2010; **26**(17): 2190–2191.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Fisher RA: **Statistical Methods for Research Workers**. Edinburgh: Oliver and Boyd, Fifth ed., 1934.  
[Reference Source](#)
- Good IJ: **Significance tests in parallel and in series**. *J Am Stat Assoc.* 1958; **53**(284): 799–813.  
[Publisher Full Text](#)
- Wilson DJ: **The harmonic mean p-value for combining dependent tests**. *Proc Natl Acad Sci U S A.* 2019; **116**(4): 1195–1200.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Dunn OJ: **Estimation of the means of dependent variables**. *Ann Math Statist.* 1958; **29**(4): 1095–1111.  
[Publisher Full Text](#)
- Šidák Z: **Rectangular confidence regions for the means of multivariate normal distributions**. *J Am Stat Assoc.* 1967; **62**(318): 626–633.  
[Publisher Full Text](#)
- Tippett LHC: **The methods of statistics**. *The Methods of Statistics*. 1931.  
[Reference Source](#)
- Hochberg Y, Tamhane AC: **Multiple Comparison Procedures**. John Wiley & Sons, New York, 1987.  
[Publisher Full Text](#)
- Marcus R, Peritz E, Gabriel KR: **On closed testing procedures with special reference to ordered analysis of variance**. *Biometrika.* 1976; **63**(3): 655–660.  
[Publisher Full Text](#)
- Vovk V, Wang R: **Combining p-values via averaging**. 2018.  
[Publisher Full Text](#)
- Brown MB: **400: A method for combining non-independent, one-sided tests of significance**. *Biometrics.* 1975; **31**(4): 987–992.  
[Publisher Full Text](#)
- Wilson DJ: **Reply to goeman et al.: Trade-offs in model averaging using multilevel tests**. *Proc Natl Acad Sci U S A.* 2019; **116**(47): 23384–23385.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Davis RA, Resnick SI: **Limit theory for bilinear processes with heavy-tailed noise**. *The Annals of Applied Probability.* 1996; **6**(4): 1191–1210.  
[Publisher Full Text](#)
- Goeman JJ, Rosenblatt JD, Nichols TE: **The harmonic mean p-value: Strong versus weak control, and the assumption of independence**. *Proc Natl Acad Sci U S A.* 2019; **116**(47): 23382–23383.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Simes RJ: **An improved Bonferroni procedure for multiple tests of significance**. *Biometrika.* 1986; **73**(3): 751–754.  
[Publisher Full Text](#)
- Uchaikin VV, Zolotarev VM: **Chance and stability: Stable distributions and their applications**. Walter de Gruyter, 1999.  
[Publisher Full Text](#)
- Nolan JP: **Parameterizations and modes of stable distributions**. *Statistics & Probability Letters.* 1998; **38**(2): 187–195.  
[Publisher Full Text](#)
- Wilks SS: **The large-sample distribution of the likelihood ratio for testing composite hypotheses**. *Ann Math Stat.* 1938; **9**(1): 60–62.  
[Publisher Full Text](#)
- Krishnaiah P, Rao M: **Remarks on a multivariate gamma distribution**. *Am Math Mon.* 1961; **68**(4): 342–346.  
[Publisher Full Text](#)
- Wilson D: **R code for Figures**. *figshare*. Software. 2020.  
<http://www.doi.org/10.6084/m9.figshare.11907033.v1>
- Rüschendorf L: **Random variables with maximum sums**. *Adv Appl*

- Probab.* 1982; **14**(3): 623–632.  
[Publisher Full Text](#)
27. Meng XL: **Posterior predictive p-values.** *Ann Stat.* 1994; **22**(3): 1142–1160.  
[Publisher Full Text](#)
  28. Sellke T, Bayarri MJ, Berger JO: **Calibration of p-values for testing precise null hypotheses.** *Am Stat.* 2001; **55**(1): 62–71.  
[Publisher Full Text](#)
  29. Neyman J, Pearson ES: **Ix. On the problem of the most efficient tests of statistical hypotheses.** *Philosophical Transactions of the Royal Society of London. Series A Containing Papers of a Mathematical or Physical Character.* 1933; **231**(694–706): 289–337.  
[Publisher Full Text](#)
  30. Held L, Ott M: **On p-values and Bayes factors.** *Annu Rev Stat Appl.* 2018; **5**: 393–419.  
[Publisher Full Text](#)
  31. Pearson K: **On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random.** *Biometrika.* 1933; **25**(3–4): 379–410.  
[Publisher Full Text](#)
  32. Liu Y, Xie J: **Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures.** *J Am Stat Assoc.* 2019; 1–18.  
[Publisher Full Text](#)

# Open Peer Review

Current Peer Review Status: 

---

Version 1

Reviewer Report 18 December 2020

<https://doi.org/10.21956/wellcomeopenres.17286.r41707>

© 2020 Rustamov R. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Raif Rustamov** 

Data Science and AI Research, Chief Data Office, AT&T Inc, Bedminster, NJ, USA

P-value combination approaches have started attracting increasing attention after the recent introduction of the Harmonic Mean P-value (HMP) <sup>1</sup> and Cauchy Combination Test <sup>2</sup> combination methods. The motivations for these two approaches are very different - Bayesian and Cauchy additive property, but interestingly, these methods are asymptotically equivalent in the most relevant regime, namely for small p-values <sup>3</sup>. A fundamental question is whether there are other p-value combination approaches that a) have a different asymptotic behavior and b) are at least as efficient as these two methods. The paper under review analyzes this problem from both theoretical and simulation aspects, and provides evidence that such a method may not exist.

The main theoretical contribution of this paper is to generalize the logic behind HMP to a parametrized class of p-value combination methods by computing the appropriate thresholds via generalized central limit theorem (GCLT). These thresholds are based on the independence assumption yet are robust to violations. They are directly useful for practical applications, and provide an alternative to the robust risk analysis (RRA) based thresholds <sup>4</sup> that are universal but rather conservative. An argument using a result from Davis-Resnick shows that there exists an explicit condition that guarantees the proper test size control at the limit. It seems to me that this condition can be relaxed; for example a theorem in <sup>2</sup> shows that a certain bivariate normality condition leads to the correct limiting size control. I take this as the evidence that the proposed thresholds can be used more widely.

It is the comparison of GCLT and RRA thresholds that is very revealing. It becomes clear that HMP, which corresponds to the case of  $r=-1$  has the most potential for gains due to the GCLT/RRA threshold ratio being  $\log K$ . For methods with  $r < -1$  the thresholds differ at most by a factor of 2 and there is little room to provide viable alternatives to HMP. Finally, methods with  $r > -1$  are neither robust to dependencies (as shown by experiments) nor provide threshold gains at the limit of large  $K$ . This can be seen as evidence to the unique place of  $r=-1$  and that the only interesting/efficient methods for p-value combination should have the same asymptotic behavior as HMP. Of course, this only establishes the asymptotic form of the tail, but the specific form can make big difference in practice.

The paper continues on to experimentally compare this class of methods using both GCLT and RRA thresholds for different values of  $r$ . The experiments consider a wide variety of setups and are rather detailed. One aspect of these experiments is that the proposed setup is very challenging due to the inclusion of all-pair dependencies, and it essentially pushes this methodology to the limit. As a consequence, the presented results may seem somewhat discouraging even for HMP. However, in my own work <sup>3,5</sup> with p-value combination approaches I have found both HMP and Cauchy Combination Test to work rather well even at the nominal rate of 0.05; the dependencies usually have limited range (e.g. correlation matrix is banded, not full) and thus the violations of independence are not as severe. Perhaps pointing out this aspect more prominently would make sure that the reader finds this methodology useful.

I would like to make a number of minor suggestions that would potentially help the readability of the paper:

1. Providing an explicit example of this new class of methods would be very helpful. For example, the combination method corresponding to  $r=-2$  marked as Levy method in Figure 6 is rather elegant and has desirable multilevel properties [personal communication with the author].
2. I find that Figure 1 and its discussion only depends on the threshold formulas not actual data. It can be moved into Section 2 to visualize the comparison of thresholds.
3. The discussion around Figure 5/excess power was not very clear to me. Perhaps a bit more detailed explanation of the notion of excess power would be helpful here.
4. Maybe include another example of data that does not have the severe dependency structure; this example does not have to be analyzed as thoroughly, but should showcase that all of these methods are practically useful.

## References

1. Wilson D: The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*. 2019; **116** (4): 1195-1200 [Publisher Full Text](#)
2. Liu Y, Xie J: Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J Am Stat Assoc*. 2020; **115** (529): 393-402 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Rustamov R, Klosowski J: Kernel mean embedding based hypothesis tests for comparing spatial point patterns. *Spatial Statistics*. 2020; **38**. [Publisher Full Text](#)
4. Vovk V, Wang R: Combining p-values via averaging. *Biometrika*. 2020; **107** (4): 791-808 [Publisher Full Text](#)
5. Rustamov R, Majumdar S: Intrinsic Sliced Wasserstein Distances for Comparing Collections of Probability Distributions on Manifolds and Graphs. *Arxiv*. 2020. [Reference Source](#)

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Machine learning and data science techniques for relational structures, graphs, networks, high-dimensional and spatial datasets with an emphasis on interpretable and statistically rigorous approaches.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

## Comments on this article

### Version 1

Author Response 02 May 2020

**Daniel Wilson**, University of Oxford, Oxford, UK

After contacting the editorial team, we decided to hold off inviting further reviewers due to the current situation. We plan to invite further reviewers once the situation has returned to normal.

**Competing Interests:** No competing interests were disclosed.

---